Ethical Considerations in NLP Shared Tasks

Carla Parra Escartín[1], **Wessel Reijers**[2], **Teresa Lynn**[2], **Joss Moorkens**[1], **Andy Way**[2] **and Chao-Hong Liu**[2]

[1]ADAPT Centre, SALIS, Dublin City University, Ireland
[2]ADAPT Centre, School of Computing, Dublin City University, Ireland
{carla.parra,wessel.reijers,teresa.lynn,joss.moorkens,andy.way,chaohong.liu}
@adaptcentre.ie

Abstract

Shared tasks are increasingly common in our field, and new challenges are suggested at almost every conference and workshop. However, as this has become an established way of pushing research forward, it is important to discuss how we researchers organise and participate in shared tasks, and make that information available to the community to allow further research improvements. In this paper, we present a number of ethical issues along with other areas of concern that are related to the competitive nature of shared tasks. As such issues could potentially impact on research ethics in the Natural Language Processing community, we also propose the development of a framework for the organisation of and participation in shared tasks that can help mitigate against these issues arising.

1 Introduction

Shared tasks are competitions to which researchers or teams of researchers submit systems that address specific, predefined challenges. The competitive nature of shared tasks arises from the publication of a system ranking in which the authors of the systems achieving the highest scores obtain public acknowledgement of their work. In this paper, we discuss a number of ethical issues and various other areas of concern that relate to the competitive nature of shared tasks. We then move to propose the creation of a common framework for shared tasks that could help in overcoming these issues.

The primary goal of shared tasks is to encourage wider international participation in solving particular tasks at hand. A second objective is to learn from the competing systems so that research can move forward from one year to the next, or to establish best practices as to how to tackle a particular challenge.

Over the past few years, the organisation of and participation in shared tasks has become more popular in Natural Language Processing (NLP), speech and image processing. In the field of NLP study, researchers now have an array of annual tasks in which they can participate. For example, several shared tasks are organised at the Conference on Natural Language Learning (CoNLL),[1] the Conference and Labs of the Evaluation Forum (CLEF),[2] or the International Workshop on Semantic Evaluation (SEMEVAL).[3] For those working on a topic that proves to be particularly challenging, it has also become a trend to propose a new shared task co-located at a related conference or workshop in order to encourage contributions from the wider community to address the problem at hand. The NLP community has seen a rapid increase in the number of shared tasks recently, with many repeated periodically while others have been organised only once. During all collocated workshops at ACL 2016 alone, a total of 9 **new** shared tasks were proposed, along with others held annually. The 2016 Conference on Machine Translation (WMT16), for instance, offered 6 shared tasks already held in previous years along with 4 new ones.[4]

A distinctive feature of shared tasks is their integral competitive nature. In the field of research ethics, the factor of competition in research projects has been shown to have potentially negative ethical consequences for upholding research

[1]http://www.signll.org/conll
[2]http://www.clef-initiative.eu/
[3]http://alt.qcri.org/semeval2016/
[4]http://www.statmt.org/wmt16/

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 66–73,
Valencia, Spain, April 4th, 2017.

integrity and the open character of scientific research. For instance, McCain (1991) has argued that increased competition for funding and publications in the field of genetics has resulted in undesirable 'secretive' behaviour of scientists, of refusal to provide access to data sets or conflicts about ownership of experimental materials. Additionally, Mumford and Helton (2001) argued that the negative perception among researchers of the intentions of their competitors might invoke unethical behaviour. These are serious consequences of elements of competition on the work of researchers. However, to date, little attention seems to have been paid to preventing these problems from arising in the organisation of shared tasks in NLP research.

With the experience gathered in our community thanks to the organisation of shared tasks over the past 30 years, we believe the time is right to initiate an open discussion on a common ethical framework for the organisation of shared tasks, in order to reduce the potential negative ethical consequences of their competitive character. Such discussions should be held in the spirit of trying to globally establish – as a research community – which ethical issues should be tackled and considered across *all* shared tasks; the purpose of this paper is not to criticise how any particular shared task is or has been organised thus far in the field.

The remainder of this paper is organised as follows: Section 2 is devoted to an overview of the role of shared tasks in NLP, including their definition, importance as well as particular issues in the existing shared tasks in our field. Section 3 is devoted to a discussion on the potential negative ethical impacts of the factor of competition that is insufficiently regulated, and finally Section 4 proposes steps towards the creation of a common framework for the organisation of shared tasks in NLP that assists at overcoming the ethical issues we identify.

2 Shared Tasks in NLP

As mentioned in Section 1, shared tasks are competitions to which researchers or teams of researchers submit systems that address a particular challenge. In the field of NLP, the first shared tasks were initiated in the United States by NIST in collaboration with DARPA (Mariani et al., 2014). Paroubek et al. (2007) report that the first shared tasks – then called evaluation campaigns – focused on speech processing and started in 1987.[5]

In 1992, new initiatives focused on the field of text understanding under the umbrella of the DARPA TIPSTER Program (Harman, 1992). Since then, researchers in NLP have experienced how this type of benchmarking for NLP tools and systems has become a tradition in many sub-areas. In fact, some of the current annual shared tasks date all the way back to 1998 and 1999 when the first SEMEVAL (then called SENSEVAL-1)[6] and CONLL[7] were organised.

Typically, shared tasks consist of 4 distinct phases (Paroubek et al., 2007):

1. Training phase,
2. Dry-run phase,
3. Evaluation phase, and
4. Adjudication phase.

During the training phase, participants are provided with data to calibrate and train their systems. Such systems are subsequently used to process a blind test set during the dry-run phase, and their results are evaluated against a 'gold standard' previously prepared by the shared task organisers. In the adjudication phase, participants are asked to raise any issues observed during the evaluation and validate the obtained results.

2.1 Why are shared tasks important in our field?

Shared tasks are important because they help boost the pace of development in our field and encourage a culture of improving upon the state-of-the-art. Shared tasks have an additional advantage: by using the same data, all systems can be evaluated objectively and comparisons across systems could be made easier.

At the same time, some best practices and *de facto* standards have evolved from shared tasks, e.g. the widely used CoNLL format used in parsing and many other NLP tasks, and the splitting of German compounds in MT proposed by Koehn and Knight (2003).

A by-product of these shared tasks are the new datasets that are made available for use by the

[5] See Pallett (2003) for an overview of these first shared tasks and the role that NIST played in them.

[6] http://www.senseval.org/

[7] http://www.cnts.ua.ac.be/conll99/npb/

wider research community. Shared tasks encourage the development of new resources and also encourage innovative approaches to data collection. Moreover, provided the data is made available, any researcher can measure the performance of their system against past shared task data. It is also possible for any researcher outside one shared task (possibly investigating different topics) to use the publicly available shared task data after the event to benchmark new systems or applications, and allow for replication of experiments, e.g. Mate Tools development reported evaluations based on datasets from the CoNLL 2009 Shared Task (Bohnet, 2010).

Shared tasks with a large number of participants can also indicate the need to tackle a particular problem, or point to challenges that are particularly attractive for the NLP research community. The participation of industry-based teams in shared tasks shows that some of them are relevant beyond the academic research community.

Taken together, shared tasks have proven themselves to be very effective in incentivising research in specialised areas, but they come at a cost: organisers need to prepare the datasets well in advance, define the evaluation criteria, gather enough interest for participation, rank the submitted systems, and so on. At the same time, there is little information sharing among shared tasks that would allow organisers to benefit from the experience of others. As a result, shared tasks vary greatly in the way they are organised, how the datasets are shared, and the type of information (and data) which is available to participants and the research community both before, during, and after the evaluation.

2.2 Variability across shared tasks

Depending on the task at hand, shared tasks are organised in different ways. In some cases (such as the MT shared tasks), no annotated data is needed, and thus only aligned bilingual data is used.

In others, prior to the shared task, the organisers create annotated data that will be distributed to all participating teams to allow them to prepare their systems for the task. Such annotated data is used with two main aims: (i) adjusting to the format required for submissions, and (ii) allowing researchers to explore the data to develop automatic systems, either rule-based or machine-learning based, so that they are able to perform the task on unseen data.

In some cases, the shared task organisers will distinguish between two different tracks for the same shared task depending on the source of the data being used to train the systems. In most cases, all teams in an evaluation test their systems on the same datasets to allow for easier across-the-board comparisons ('closed' track). Other shared tasks allow for the inclusion of additional data by individual teams ('open' track). In these 'open' tracks, the inclusion of other data is not necessarily verified and based on a trust system. It is worth noting that, to date, this system has worked well and, to the best of our knowledge, there have been no known issues of mistrust in NLP shared tasks.

Depending on the type of shared tasks, different evaluation methodologies will be used, ranging from purely automatic metrics, such as precision and recall for many of the shared tasks focusing on Information Retrieval, to human evaluation, such as the ranking of MT outputs or automatically generated text.[8]

3 Potential ethical issues concerning the competitive nature of shared tasks

As we have seen in the previous section, there is currently a great variability and a lack of standardisation in the organisation of shared tasks. Because shared tasks have become an important part of the scientific research in NLP, a certain level of standardisation is nonetheless required in order to safeguard satisfactory levels of scientific integrity and openness of scientific research. This standardisation in the organisation of shared tasks is needed specifically to mitigate potential negative ethical impacts of their competitive character. With a view to proposing a standard approach to shared task organisation, in this section, we discuss the potential negative ethical impacts of competition in scientific research and subsequently illustrate this by addressing potentially problematic aspects of the organisation of shared tasks in NLP.

3.1 Ethical issues arising from competition in scientific research

Competition is a factor in scientific research that is not limited to the field of NLP. In the organisation of contemporary science, competitive schemes for

[8] Paroubek et al. (2007) offer a good overview of the organisation of shared tasks and the different types of evaluation that one may come across.

scientific positions, publication possibilities or research funding are increasingly influential. However, shared tasks are distinctive from traditional forms of research, such as the writing of individual research papers or the organisation of experiments in a closed research project, because the element of competition is integral to the research activity. In other words, shared tasks not only take place within a competitive context, they are competitions *per se*.

For this reason, the effects of competition on the conduct of researchers should be taken seriously. As Anderson et al. (2007, p. 439) argue: "the relationship between competition and academic misconduct is a serious concern". A number of negative ethical impacts of competition in scientific research are discussed in the literature on research ethics. We suggest that the NLP community could draw on previous experiences and studies in the wider scientific community. We present three of the most important ones:

- **Secretive behaviour.** This effect of competition results from the tendency of researchers to give themselves an unfair competitive advantage in terms of knowledge concerning the research challenge at hand. McCain (1991) suggests that this behaviour can have several concrete forms, such as the unwillingness to publish research results in a timely fashion, refusal to provide access to data sets and conflicts concerning the 'ownership' of experimental materials.

- **Overlooking the relevance of ethical concerns.** Another effect of competition is the tendency of the teams competing to overlook the relevance of ethical concerns in their research. As Mumford and Helton (2001) explain, this might have the form of disregarding ethical concerns in general, or specifically with regard to one's own work while anticipating the potential ethical misconduct of others ("if they can do it, why shouldn't we?"). This can lead to careless – or questionable – research conduct.

- **Relations with other scientists.** Because the stakes in competitions can be very high (they might result in further or decreased research funding, or in opening up or closing off of future career paths), competitions might have negative impacts on the relations between peers (Anderson et al., 2007). This might lead researchers to have the tendency to behave unethically with regards to their peers in order to preserve or strengthen their reputation.

3.2 Potential negative effects of competition in shared tasks in NLP

The motivation for involvement in shared tasks has evolved somewhat over the recent past. Many researchers in MT, for example, will participate in the annual shared tasks organised at WMT, where there is a ranking of the best systems for a proposed task. Participation and success in tasks such as these are often used to demonstrate research excellence to funding agencies. At the same time, performance of the systems may also have a greater impact on the funding for a complete research area (not only for individual teams or institutions). We only need to look back to the notorious ALPAC report (Pierce and Carroll, 1966), whose consequences for research funding in the US for MT were devastating for a considerable period. Such funding-related motivation can in turn lead to increased competitiveness.

When we revisit the shared tasks within NLP, the potential negative ethical impacts of competition identified in the literature on research ethics can also be found in this field. Here, we discuss the main issues identified which require mechanisms to be established by our community to prevent them from happening.

- **Secretiveness.** Competitiveness can sometimes lead to secretiveness with respect to the specific features used to tune a system to ensure that the best methods and/or approaches stay in the same institution/team. Participants usually submit their system descriptions to the shared task, in the form of presentations and research papers. However, the way in which such systems are described may vary greatly, as one can always choose a more abstract higher-level description to avoid 'spilling the beans' about the methodology applied, and retaining the knowledge rather than sharing it.

- **Unconscious overlooking of ethical concerns.** Leading on from the secretiveness issue raised above, teams may unintentionally be vague in reporting details of their systems'

parameters and functionality solely on the basis that other teams have also previously reported in this way. Such practice can simply arise from the existence of convention in the absence of guidelines or standards.

- **Potential conflicts of interest.** Finally, another potential ethical issue is related to organisers or annotators being allowed to participate in the shared task in which they are involved. Again, while some find it unethical to participate in their own shared task, others disagree, and the community trusts that in such cases the organisers trained their systems under the same conditions as the rest of the participants, i.e. they did not take advantage of prior access to data and did not train their systems for a longer period of time, or have a sneak peak at – or hand-select for optimal performance – the test data to improve their system's performance and cause it to be highly ranked. Both points of view are perfectly valid, and in some cases even justified, e.g. teams working on (usually low-resourced) languages for which they themselves are one of the few potential participants for those languages. While the overlap of organisers, annotators and participants has not yet revealed itself to be a major issue in our field, and the goodwill and ethical conduct of all involved is generally trusted, it is worth considering the establishment of methods for minimising the risk of this happening in the future. One such measure could be for the organisers to explicitly state whether the overlap is likely to happen.[9]

Subsequently, we have identified a number of other potential conflicts with the objectivity and integrity of research that may arise from the competitive nature of shared tasks in NLP. Whether intentional or unintentional, these issues are worth considering when developing a common framework for the organisation of and participation in shared tasks:

- **Lack of description of negative results.** The fact that negative results are also informative is something that no researcher will deny. However, as shown by Fanelli (2010), researchers have a tendency to report only positive results. He claims that this may be because they "attract more interest and are cited more often", adding that there is a belief that "journal editors and peer reviewers might tend to favour them, which will further increase the desirability of a positive outcome to researchers".

  Furthermore, with PhD students being pressed to publish in the top conferences in their fields, they may be reluctant to submit systems that do not report on positive results. As a result, while we always discover what worked for a particular task, we are not usually told what did *not* work, although that may be of equal or (even) greater importance than the methodology that worked, as it would help others to avoid repeating the same mistake in the future. In fact, it may be the case that the same approach has been tested by different institutions with no success and that we are incurring a hidden redundancy that does not help us to move forward as a field. In order to prevent these issues from occurring, we should design mechanisms that incentivise the publication of negative results and more thorough error analysis of systems

  Similarly, it may be the case that although it is highly desirable that industry-based teams participate in a shared task, some may be reluctant to do so on the basis of the negative impact that this may have for their product if it does not end up among the first ranked. Thus, rather than strengthening the academia-industry relationship and learning from each other, we risk making the gap between the two bigger rather than bridging it. Should we not address this and establish mechanisms that encourage industrial teams to participate in shared tasks without such associated risks?

- **Withdrawal from competition.** Some teams may prefer to withdraw from the competition rather than participate if they fear that their performance may have a negative impact in their future funding: how could research excellence on a particular topic be argued if one's team came last in a competition? Again, mechanisms could be de-

[9] This type of overlap was highlighted by the organisers of the PARSEME shared task at the 13th Workshop on Multiword Expressions (MWE 2017): `http://bit.ly/2jPsu2n`.

signed with the aim of discouraging this type of withdrawal. For example, one possible solution would be to only report on the upper 50% of the ranked systems.

- **Potential 'gaming the system'.** Another concern is the impact of the results of the shared task beyond the shared task itself (e.g. real-world applications, end-users). Shared tasks are evaluated against a common test set under the auspices of a 'fair' comparison among systems. However, as the ultimate goal of most participating teams is to obtain the highest positions in the ranking, there is a risk of focusing on winning, rather than on the task itself. Of course, accurate evaluation is crucial when reporting results of NLP tasks (e.g. Summarisation (Mackie et al., 2014); MT (Graham, 2015)). As evaluation metrics play a crucial role in determining who is the winner of a shared task, many participating teams will tune their systems so that they achieve the highest possible score for the objective function at hand, as opposed to focusing on whether this approach is actually the best way to solve the problem. This, in turn, impacts directly on the real-world applications for which solving that challenge is particularly relevant, as it may be the case that the 'winning' systems are not necessarily the best ones to be used in practice.

  As discussed previously, some shared tasks allow for 'closed' and 'open' variants, i.e. in the 'closed' sub-task, participants use only the data provided by the shared task organisers, such that the playing field really is level (we ignore for now the question as to whether the leading system really is the 'best' system for the task at hand, or (merely) has the best pre-preprocessing component, for instance). By contrast, in the 'open' challenge, teams are permitted to add extra data such that true comparison of the merits of the competing systems is much harder to bring about.

- **Redundancy and replicability in the field.** Another important issue is that, although this should be the overriding goal, we typically find that for any new data set – even for the same language pair – optimal parameter settings established in a previous shared task do not necessarily carry over to the new, albeit related challenge. This is a *real* problem, as if this is the case, we should ask ourselves what we as a field are really learning. At the same time, our field experiences a lot of redundancy, as we try to reimplement others' algorithms against which we test our own systems. This is the case particularly when systems participating in a shared task are not subsequently released to the community.[10]

- **Unequal playing field.** Another potential risk is the fact that larger teams at institutions with greater processing power (e.g. better funded research centres or large multinationals) may have a clear unfair advantage in developing better performing systems, rendering the 'competition' as an unequal playing field for researchers in general. This could be mitigated against by establishing, beforehand, the conditions under which systems are trained and tested for the task.

In this section, we have identified several potential ethical concerns related to the organization and participation in shared tasks. As observed, the three issues discussed in the academic literature on competition in research (cf. Section 3.1) appear to be important considerations for shared tasks in NLP. In addition, we have highlighted some other areas of potential ethical consideration in our field with respect to shared tasks. In the next section, we discuss potential paths to tackle the ethical concerns raised here.

4 Future directions

The great value of shared tasks is there for all to see, and there is no doubt that they will continue to be a major venue for many researchers in NLP in the future. Nonetheless, we have pointed out several ethical concerns that we believe should be addressed by the NLP community, and mechanisms created to prevent them should be also agreed upon. At the same time, there may be other ethical considerations that the authors have omitted due to lack of knowledge about *all* shared tasks

[10] The existence of initiatives such as CLARIN[11] or the recent efforts made by ELDA to try to standardize even various versions of the 'same' dataset, evaluation metric, or even a particular run of an experiment show that we are shifting to a new research paradigm where open data, research transparency, reproducibility of results and a collaborative approach to advancements in science are advocated (Pedersen, 2008; Perovšek et al., 2015).

in NLP, or simply because they arose within participation in specific shared tasks and have never been shared with the community. Thus, we see that a first step towards determining potential ethical issues related to the organisation of and participation in shared tasks is to conduct a survey in our community to ensure broad contribution. Such a survey – to be launched shortly after discussions at the 2017 Ethics in NLP workshop – consists of two parts. The first tries to gauge the varying requirements of shared tasks, and the second one aims at assessing what people feel are important factors for consideration when drawing up a common framework for shared tasks in NLP. This common framework will ensure greater transparency and understanding of shared tasks in our community, and prevent us from encountering the potential negative impact of the ethical concerns raised here.

Questions regarding past experiences related to shared tasks (either as organisers, annotators or participants) are included in the survey to gather information regarding (i) best practices used in specific shared tasks that could be extrapolated to new ones, (ii) the type of information that is available to participants before, during and after the shared task, (iii) potential ethical concerns encountered in the past and how they were tackled, (iv) other causes for concern from the NLP community and (v) good experiences that we should aim at replicating.

Besides recommendations on best practice, we envisage the creation of shared task checklists based on the questions in the survey and their replies. These checklists would target the organisers, annotators and participating teams in shared tasks, and would be used to state any relevant information required in each case. By subsequently making them publicly available to the community (e.g. at the shared task website), any participating team or researcher interested in the shared task topic would know how specific topics were addressed in the shared task, and what information was or will be available to them. What follows is a non-exhaustive list of some of the items that we foresee including in the checklist (subject to discussion and amendment):

- Participation of organisers in the shared task;
- Participation of annotators or people who had prior access to the data in the shared task;
- Public release of the results of the participating systems after the shared task, under an agreed license;
- Declaration of the list of contributors to a certain system at submission time;
- Anonymisation of the lower (50% ?) of systems evaluated to be referred to by name in published results;
- ...

5 Conclusion

In this paper we have discussed a number of potential ethical issues in the organisation and participation of shared tasks that NLP scientists should address to prevent them from arising as problems in the future. Besides taking into account the particular features of shared tasks, we investigated the potential ethical issues of competition in scientific research and extrapolated such issues to the potential problems that may arise in our own field. In addition, as we believe this should be tackled by the NLP community as a whole, we have proposed the launch of a survey to gather further information about shared tasks in NLP that will help in the development of a common framework in the near future. This would include current best practice, a series of recommendations and checklists as to what issues should be taken into account, as well as what information is provided to participants, depending on the type of shared tasks in question.

Finally, shared tasks in our field play an essential role in NLP. They have undoubtedly helped improve the quality of the systems we develop across a range of NLP sub-fields, to a point where many of them comprise essential components of professional workflows. The system as such is not irretrievably broken, so there may be a temptation to not fix the issues outlined in this paper. However, we firmly believe that the field of NLP has reached a level of maturity where some reflection on the practices that we currently take for granted is merited, such that our shared tasks become ever more reliable and consistent across our discipline, and further strides are made to the benefit of the field as a whole as well as to the wider community.

Acknowledgements

The authors wish to thank the anonymous reviewers for their valuable feedback. This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

References

Melissa S. Anderson, Emily A. Ronning, Raymond de Vries, and Brian C. Martinson. 2007. The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, 13(4):437–461.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniele Fanelli. 2010. Do pressures to publish increase scientists' bias? an empirical support from us states data. *PLoS ONE*, 5(4):e10271, 04.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China, July. Association for Computational Linguistics.

Donna Harman. 1992. The darpa tipster project. *SIGIR Forum*, 26(2):26–28, October.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound splitting. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stuart Mackie, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2014. On choosing an effective automatic evaluation metric for microblog summarisation. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 115–124. ACM.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, and Olivier Hamon. 2014. Rediscovering 15 years of discoveries in language resources and evaluation: The lrec anthology analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Katherine W. McCain. 1991. Communication, competition, and secrecy: The production and dissemination of research-related information in genetics. *Science, Technology & Human Values*, 16(4):491–516.

Michael D. Mumford and Whitney B. Helton. 2001. Organizational influences on scientific integrity. In Nicholas Steneck and Mary Scheetz, editors, *Investigating research integrity: Proceedings of the first ORI research conference on research integrity*, volume 5583, pages 73–90.

David S. Pallett. 2003. A look at nist's benchmark asr tests: past, present, and future. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 483–488, Nov.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of Evaluation in Natural Language Processing. *Traitement Automatique des Langues*, 48(1):7–31, May.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Matic Perovšek, Vid Podpečan, Janez Kranjc, Tomaž Erjavec, Senja Pollak, Quynh Ngoc Thi Do, Xiao Liu, Cameron Smith, Mark Cavazza, and Nada Lavrač. 2015. Text mining platform for nlp workflow design, replication and reuse. In *Proceedings of the Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015*.

John R. Pierce and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, Washington, DC, USA.